

Analysis on Newcomb's paradox, free will and rationality

SAYAK SAMANTA

Newcomb's paradox is a paradoxical thought experiment by William Newcomb. The experiment is as follows.

In a room, there are two boxes. One of the boxes contains \$1000 while the other is a *mystery box*. Along with the boxes, there is also a *super-computer* in the room. This highly intelligent computer has a remarkable ability to predict people and has made correct predictions of many people in the past. Along with this fact, you have the choice to either take the \$1000 and the mystery box or only the mystery box. However, before you came inside the room or even knew about the experiment, the computer had made a prediction as to what you'd choose.

If the computer predicted you chose to take the box with \$1000 and the mystery box, then the supercomputer put nothing in the mystery box. If the computer predicted you would choose only the mystery box, then it put \$1,000,000 in the mystery box.

Keep in mind that the supercomputer has already made its prediction and set up the boxes with the money before you even entered the room. Its only goal is to make the correct prediction, and your goal, you being a rational human, is to take the most money. Do you take both boxes or only the mystery box?

The paradox nearly splits people 50/50 as to which is the correct answer. We can classify these people into two groups—the one-boxers and two-boxers. Strangely, both sides swear their answer is correct and the opposition is ignorant. Even the smartest of people are divided on this problem. So, who is correct?

The One-boxer

To understand who is correct, we must look into both sides and see their claims.

One-boxers say that since this supercomputer has made correct predictions of many people in the past, it is reasonable to believe that it will correctly predict your

choice. If you chose two boxes, it would have predicted that and put nothing in the mystery box. However, if you chose only the mystery box, then the supercomputer would have put \$1,000,000 in the mystery box as it would have predicted your decision beforehand. Therefore, choosing only the mystery box returns the most money. If thinking is not your game, let us prove this mathematically using probabilities.

1. Let us say that the chance that the super-computer is correct is "C" and it is not correct, "1 - C".

C = probability prediction is RIGHT

1 - C = probability prediction is WRONG

Now, let us see the Expected Utility or E.U. of one-boxing or two-boxing. The E.U. will show us what is the average money we can expect to make from either choice.

(Computer is)	RIGHT	WRONG
One box	\$1,000,000 • C	0 • (1 - C)
Two box	\$1,000 • C	\$1,001,000 • (1 - C)

$$\begin{aligned}
 \text{(One-box) E.U.} &= \$1,000,000 \bullet C + 0 \bullet (1 - C) \\
 &= \$1,000,000 \bullet C
 \end{aligned}$$

$$\begin{aligned}
 \text{(Two-box) E.U.} &= \$1,000 \bullet C + 1,001,000 \bullet (1 - C) \\
 &= \$1,001,000 - \$1,000,000 \bullet C
 \end{aligned}$$

If these two equations are equated, we will get the C value for which they are both the same. This value is 50.05%. Therefore, if the computer is better than guessing at random, we can expect a higher return from one-boxing. Since we know, that this computer is very good in predicting decisions, and that it has predicted correctly a lot of times, it is very reasonable to one-box.

The Two-boxer

By now, I assume most of you are convinced of being a one-boxer. However, in the words of Robert Nozick— *to almost everyone it is perfectly clear and obvious what should be done; the difficulty is that these people seem to divide almost evenly on the problem with large numbers thinking the opposite half is just being silly.*

In 2016, a newspaper polled over 31,000 people about this problem, 53.5% were one-boxers while 46.5% were two-boxers. Let us see the school of thought of the two-boxers.

Two-boxers say that since the super-computer has made its prediction before you entered the room, whatever decision you make will not change what's in the mystery box. Whether you two-box or one-box, it will not change whether there is a \$1,000,000 or \$0 in the mystery box.

If there is \$0 in the mystery box, one-boxing gives us \$0 while two-boxing gives us \$1,000. If there is a \$1,000,000 in the mystery box, one-boxing would give us \$1,000,00 and two-boxing returns us \$1,001,000.

Mystery Box	One-boxing	Two-boxing
\$0	\$0	\$1,000
\$1,000,000	\$1,000,000	\$1,001,000

As you can see from the table above, you are always better off by picking two-boxes (\$1,000 or \$1,001,000). This is known as *Strategic Dominance* where you always pick the dominant strategy which in this case is two-boxing.

Who's Right?

Although both arguments seem logical, there is a hidden assumption in each that divides us evenly. What exactly is this supercomputer?

Does this supercomputer use history and past experiences to predict the future or is it an all-knowing computer that uses a good algorithm to predict the future with extreme accuracy? In the second case, it is rational and reasonable to go one-boxing as the supercomputer would know you'd do that. However, in the first case, it is also reasonable to go two-boxing.

Regardless, both one-boxers and two-boxers agree that it is rational to one-box if there are multiple rounds as the supercomputer would learn from previous rounds that you would one-box.

Let us say we replace the supercomputer with a magical sorcerer who can predict the future. In this case, we would see a larger majority of people one-boxing. So it seems that most people are confused by what it means when the supercomputer predicts the future. Does it actually know the future or is it using history? Most people

assume it does not know the future as that would be impossible (as far as quantum mechanics goes). This assumption is what mainly divides people.

Although this may seem like a meaningless problem or something to debate on, it reveals a lot about ourselves and the world around us, as we will see.

What is free will and what does it mean to be rational?

Newcomb's paradox underlines an important fact that most of us may not have thought of. Do we have free will? By free will, I mean do we have the power to change our decisions or change the future? If this supercomputer can predict the future with extreme accuracy or even know the definite future, then do we have a choice? Some say free will is an illusion. That free will does not exist but we have to pretend it does for our society to function properly.

If a criminal is convicted of some crimes, one could argue he did not have free will and he could be given community service instead of a harsh prison sentence. But if this were to happen, people would realise crimes did not come with serious consequences, leading to chaos within society.

Another point to consider is that if we don't have free will, does it change what we do?

Apart from free will, another question has come to arise: What does it mean to be rational?

In this experiment, two-boxers pose a rational argument: the computer has already set up the boxes with the money and my decision cannot change the past. However, in the words of David Lewis, one-boxers taunt them by saying, "If you're so smart, why ain'cha rich?"

According to his paper, one-boxers end up with more money because of the rewarded irrationality of the game. The experiment rewards predicted irrationality richly. The one-boxers think that two-boxers are not rich because they have irrationally chosen not to have their millions.

The two-boxers reply by saying they were never given any choice as to whether to have a million. The reason why we (two-boxers) are not rich is that the riches are reserved for the irrational.

The two-boxers' argument is the most rational. However, this rationality will not make you rich.

This experiment makes us redefine rationality. Most of the time, a rational person does rational acts. But sometimes, for someone to be rational, they must do an irrational act. You may expect a rational society to be made up of rational people. However, as said previously, a rational person sometimes must be irrational to remain rational. Therefore, it can be said that rationality at different levels is incompatible.

Another thing that one-boxers and two-boxers have in common, is if they could pre-commit to the computer. Both sides would commit to one-boxing which returns them the most money. The super-computer would know that making the most money is the goal of a rational human, and would predict that they would one-box and put a \$1,000,000 in the mystery box. However, there are scenarios where staying true to ones word does not pay off.

Mutually Assured Destruction

On the 29th of August, 1949, the Soviet Union detonated their first bomb, the RDS-1, as part of their nuclear weapons test. This sent the U.S. and Soviet Union into a nuclear arms race. By 1960, the U.S. had 30,000 nuclear warheads while the Soviet Union had 6,000. Both sides were more than capable of destroying each other.

The idea was that if you attack us, we will attack you, in hopes that it would deter either side from starting a nuclear war. This became known as Mutually Assured Destruction or MAD.

Now say, hypothetically, you were the president of the U.S. and you just received word that the Soviets have launched their missiles, and when you were elected you pre-committed to retaliation. Do you launch all your missiles at them leading to a total annihilation of both sides and at worst a nuclear winter that kills everyone on the planet, or do you instead let them annihilate you? This is comparable to the experiment because before the experiment you could pre-commit to the supercomputer that you would one-box, but when given the choice you should instead two-box, ideally giving you \$1,001,000.

In such a situation of the Cold War, you would want a president who appears to want to retaliate but secretly would not actually launch retaliatory strikes. This comes with a risk that is anyone finds out if wouldn't retaliate, you could be thrown out of office. The important idea to deter a nuclear war than actually retaliate.

I conclude by saying Newcomb's paradox is more than a simple thought experiment to debate about. It makes us reconsider what rationality means, and the idea of free will.

Credits – [Veritasium \(Derek Muller\)](#); [Why Ain'Cha Rich \(David Lewis\)](#)